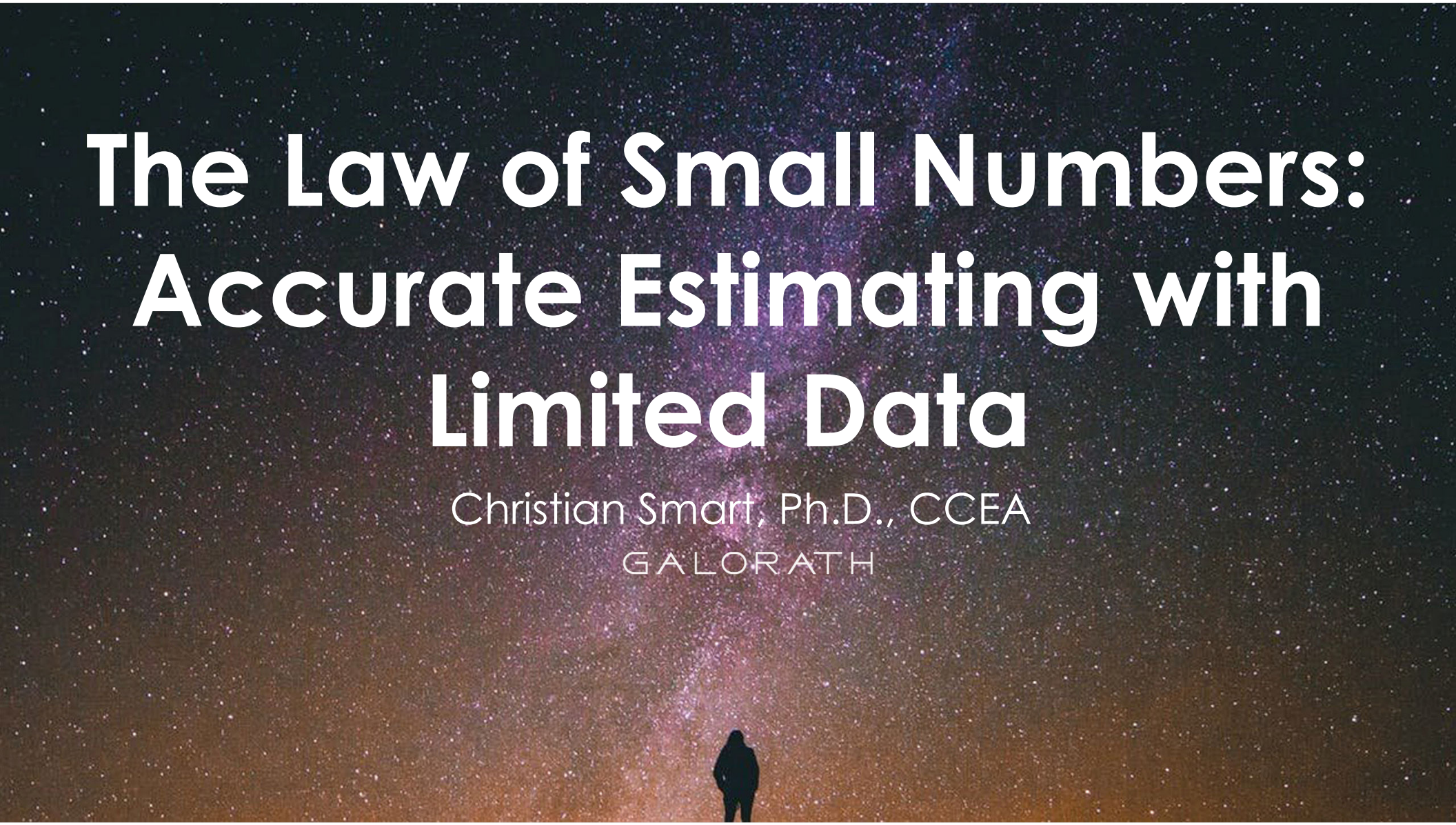


The Law of Small Numbers: Accurate Estimating with Limited Data

Christian Smart, Ph.D., CCEA
GALORATH

EXPERIMENT

- **A DRUG HAS AN 80% SUCCESS RATE**
- **FIVE PEOPLE ARE GIVEN THE DRUG**
- **WHAT IS THE PROBABILITY THAT EXACTLY FOUR OUT OF THE FIVE ARE CURED?**

GALORATH

The Law of Large Numbers

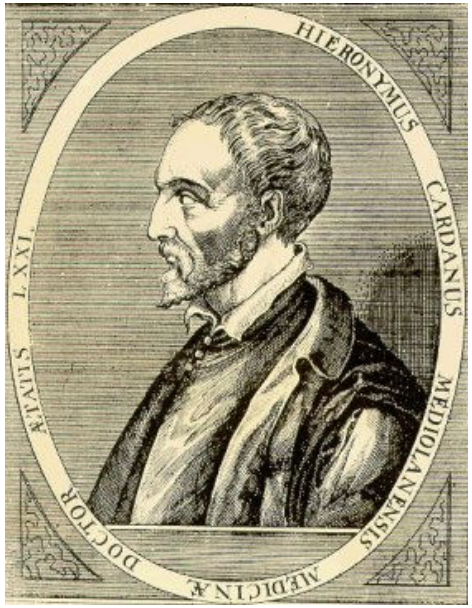
STANDARD STATISTICS REQUIRES LARGE DATA SETS

The Law of Large Number is a foundational concept in statistics:

If a sample is large enough, the sample average should be close to the mean

- Cardano proposed the concept

- Jacob Bernoulli established the first mathematical theorem



Gerolamo Cardano

Jacob Bernoulli

The key results of classical statistics require large data sets.

GALORATH

THE LAW OF SMALL NUMBERS

ISSUES WITH SMALL DATA SETS

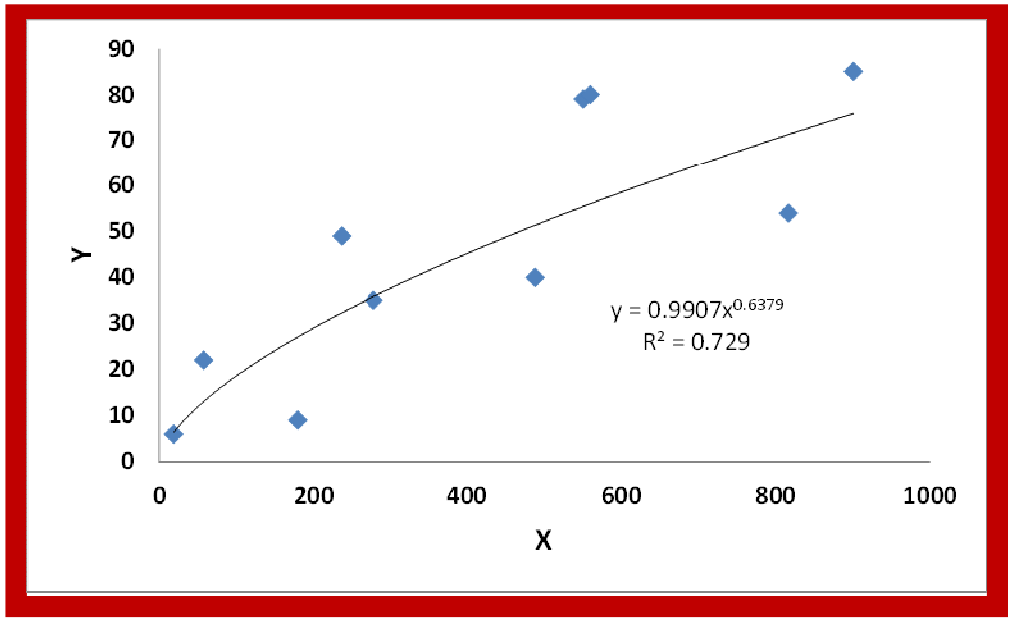
For several decades the psychologists Daniel Kahneman and Amos Tversky studied biases that affect our decision-making processes. In 2002, their research was honored with a Nobel Prize in Economics. One of their key contributions was a paper "Belief in the Law of Small Numbers." In this paper, Kahneman and Tversky define the Law of Small Numbers as the mistaken belief that a small sample accurately reflects the probabilities of a population.

In small data sets you can find patterns where none exist!

GALORATH

FOOLED BY RANDOMNESS

EXAMPLE 1

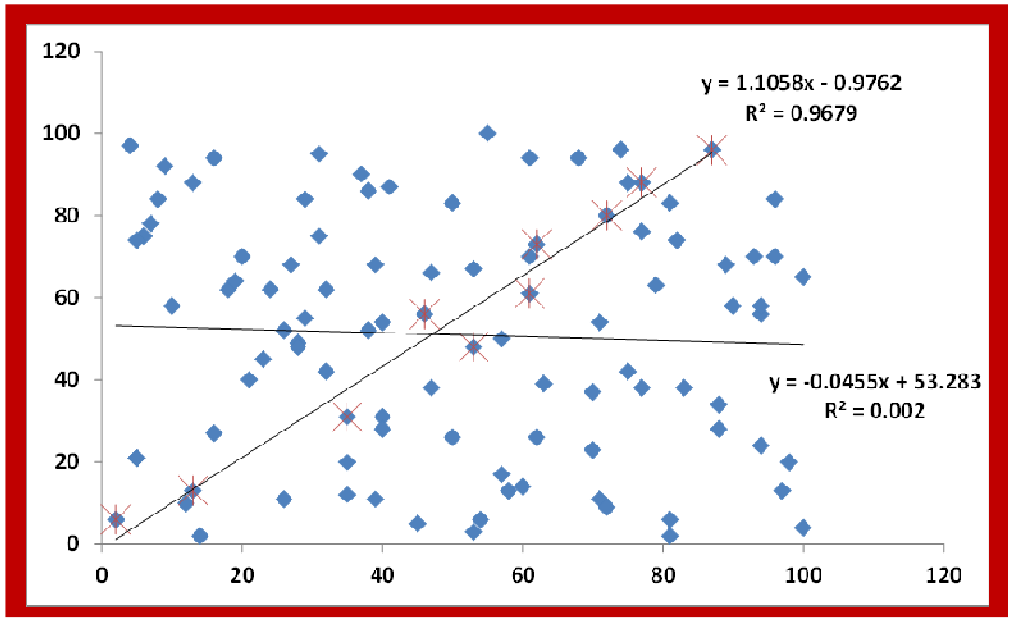


- **CONSIDER THE DATA SET ON THE LEFT**
- **ONLY 10 DATA POINTS BUT THERE IS CLEARLY A STRONG TREND!**
- **ACTUALLY, THE DATA WERE RANDOMLY GENERATED!**

GALORATH

FOOLED BY RANDOMNESS

EXAMPLE 2



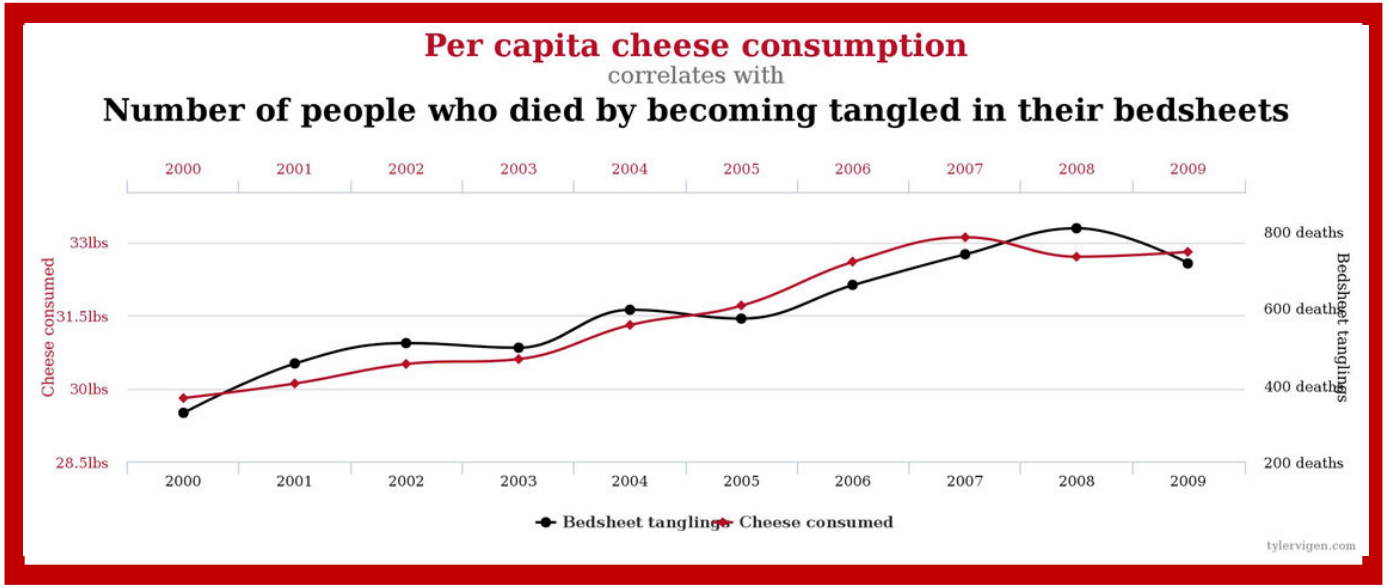
- **IN A POPULATION OF 100 DATA POINTS THAT HAVE NO TREND IT IS POSSIBLE TO FIND SMALL SAMPLES THAT DISPLAY A TREND**
- **THE DATA POINTS IN THE POPULATION HAVE ZERO CORRELATION**
- **BUT THERE ARE SAMPLES OF 10 DATA POINTS THAT HAVE STRONG TRENDS**

GALORATH

FOOLED BY RANDOMNESS

EXAMPLE 3

- **IT IS EASY TO FIND EXAMPLES OF SPURIOUS CORRELATIONS IN SMALL DATA SETS**
- **TYLER VIGEN HAS DEVOTED A WEBSITE AND A BOOK TO THE SUBJECT**



GALORATH

DATA – HOW SMALL IS SMALL?

FITTING TO NOISE

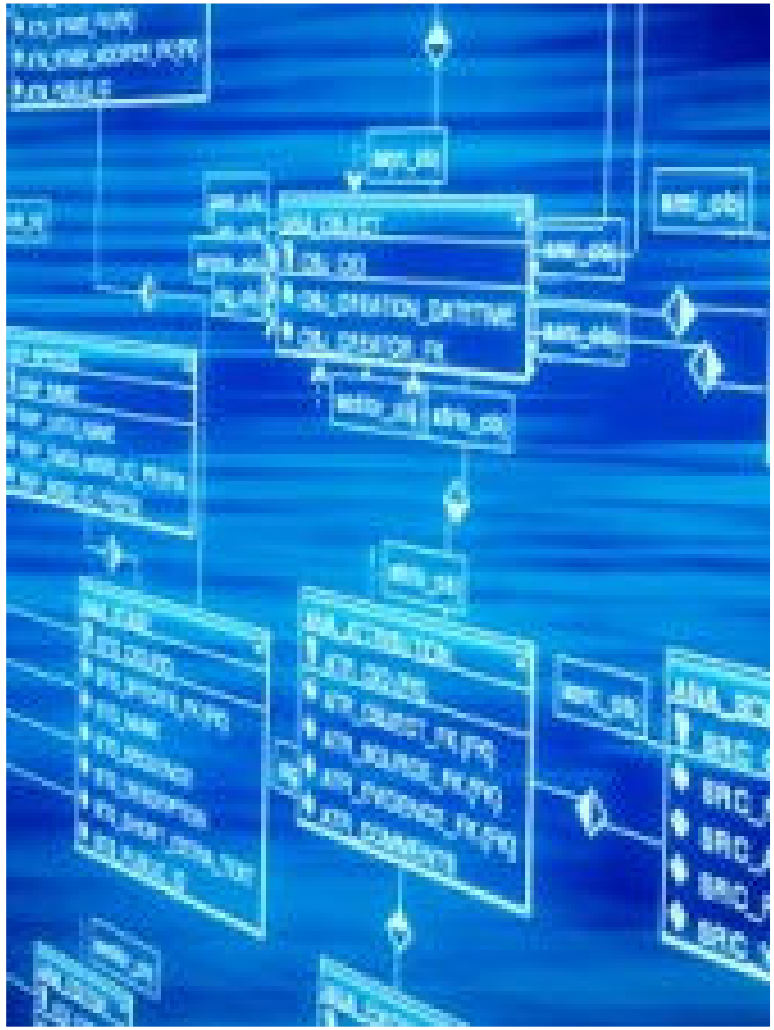
STUDIES INDICATE THAT 30-70% OF VARIABLES INCLUDED IN STEPWISE REGRESSION ARE PURE NOISE

RULE OF THUMB

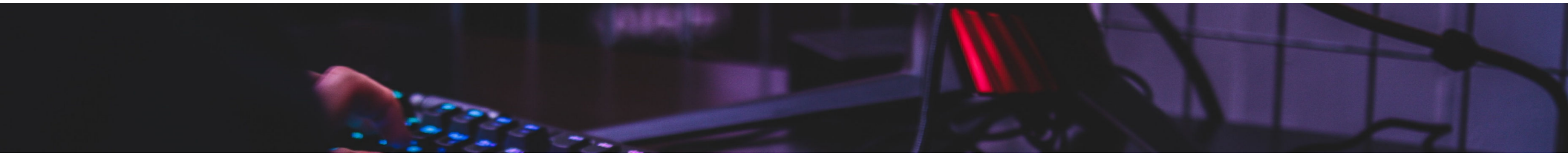
50 DATA POINTS PLUS 10 DATA POINTS PLUS 10 DATA POINTS FOR EVERY ADDITIONAL PARAMETER

BASIS

THE RULE OF THUMB IS BASED ON SIMULATION STUDIES OF FITTING REGRESSIONS TO RANDOMLY GENERATED DATA



GALORATH



WHAT CAN BE DONE?

COLLECT MORE DATA

WORK WITH OTHER GOV'T AGENCIES (E.G., AIR FORCE, COMMERCIAL COMPANIES)

CAN BE DIFFICULT TO DO

IMPUTE

USE IMPUTATION TO FILL IN MISSING DATA

NEED TO BE CAREFUL TO CONSERVE CORRELATION STRUCTURES

GO LOWER

ESTIMATE AT THE COMPONENT LEVEL INSTEAD OF THE SUBSYSTEM LEVEL

MORE NOISE

BAYES

USE TECHNIQUES DESIGNED TO ESTIMATE WITH LIMITED DATA

SUBJECTIVE

SHOW ME THE DATA!

COLLECT MORE DATA

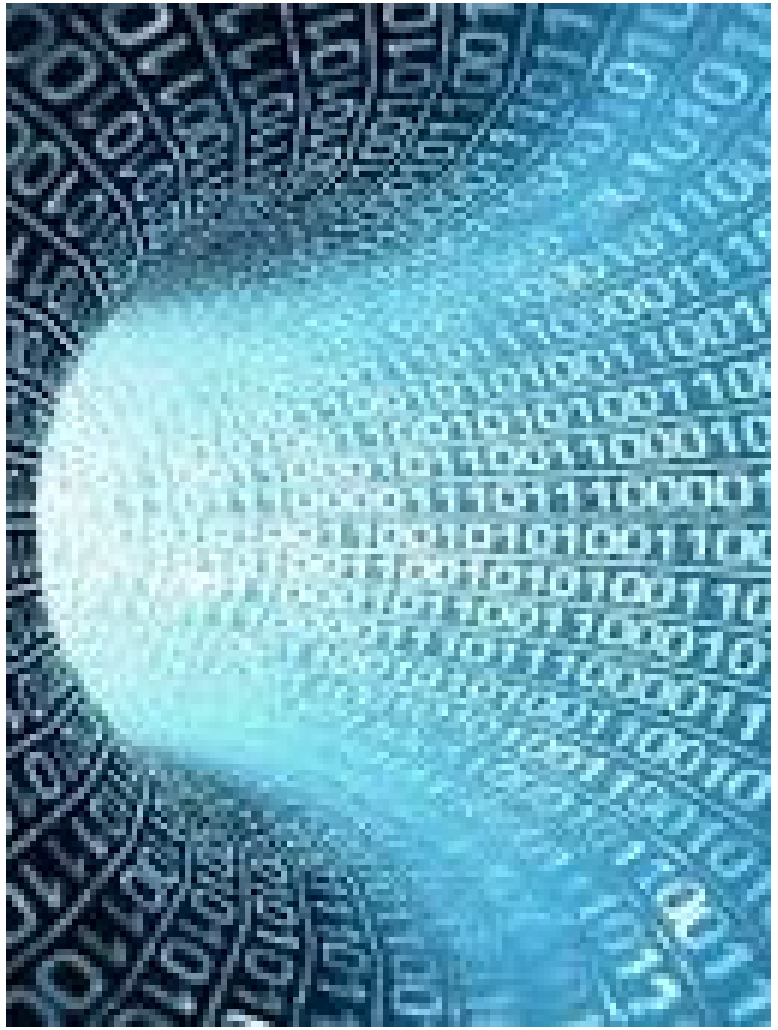
THERE ARE AIR FORCE AND COMMERCIAL MISSIONS WHOSE DATA WOULD BE USEFUL TO NASA

STATUS QUO

NASA HAS ALREADY DONE SIGNIFICANT WORK IN THIS AREA WITH THE ESTABLISHMENT AND COLLECTION OF CADREs

CAVEATS

LOW HANGING FRUIT HAS BEEN COLLECTED, THIS OPTION REQUIRES MORE EFFORT



GALORATH

IMPUTE!

MISSING DATA

- EVEN WITH SIGNIFICANT DATA SETS, THERE WILL BE MISSING FIELDS
- CUTS DOWN ON THE NUMBER OF DATA POINTS THAT CAN BE USED FOR MODELING
- IMPUTATION USES THE DATA YOU HAVE TO FILL IN MISSING INPUTS

STATUS QUO

- NOTHING SIGNIFICANT HAS BEEN DONE IN THIS AREA

POTENTIAL

- THIS IS A LOW-INTENSITY EFFORT – TIME AND RESOURCES
- ONE OF THE EASIEST OPTIONS

GALORATH

GO LOWER

ESTIMATE AT A LOWER LEVEL

- BREAK DOWN UNIQUE SUBSYSTEMS TO COMPONENTS THAT HAVE COMMONALITY WITH OTHER MISSIONS
- INCREASES NUMBER OF AVAILABLE DATA POINTS WHEN UNIQUE MISSIONS ARE MODELED

STATUS QUO

- MANY MODELS ARE AT THE SUBSYSTEM LEVEL – PCEC, NICM, EVEN SEER-SPACE

POTENTIAL

- WILL REQUIRE NEW APPROACH TO DATA COLLECTION – CADREs ARE TYPICALLY AT THE SUBSYSTEM LEVEL
- CONSIDER DEFENSE DEPT.'s FLEXFILE APPROACH
- SOME COMMERCIAL MODELS ARE AT THE LOWER LEVEL (E.G. SEER-H)

GALORATH

BAYESIAN PARAMETRICS

USE METHODS DESIGNED FOR SMALL

DATA SETS

THE BAYESIAN APPROACH TO PARAMETRICS ALLOWS YOU TO LEVERAGE ALL AVAILABEL DATA INCLUDING EXPERIENCE

STATUS QUO

VERY LIMITED USE IN PRACTICE

POTENTIAL

REQUIRES NO ADDITIONAL COLLECTION EFFORTS

REQUIRES A MODERATATE AMOUNT OF EFFORT TO CORRECTLY APPLY

The Rev. Thomas Bayes

Pierre-Simon Laplace

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

GALORATH

Imputation

MISSING DATA

Dealing with holes in your data set

When developing multivariate CERs, you may initially have a decent-sized data set

Then you notice a few missions are missing one key independent variable, a few others are missing another, etc.

Pretty soon you have whittled down your data set to only a few missions that have all the independent variables you want to consider

One option would be to go out and spend time and effort on collecting the missing information

But what if you could use statistics of the data set to make reasonable assessments of the missing values?

It turns out you can with statistical imputation

IMPUTATION

Filling in holes in your data

Well-developed statistical theory

Methods have been developed for a variety of missing types, whether at random or not at random

Important to use techniques that do not change the correlation structure of the model (if intent is to use the data for multiple regression)

Imputation techniques use a combination of maximum likelihood methods along with Bayesian techniques

Bayesian Parametrics

BAYES' THEOREM

CONDITIONAL PROBABILITY

NUMEROUS APPLICATIONS

1

ENIGMA

Bayesian techniques were used to help crack the Enigma code in Word War II, shortening the war

2

PROPERTY and CASUALTY INSURANCE

Used for over a century to set premiums when there is limited data

3

HEDGE FUND MANAGEMENT

Also used in election forecasting and game theory

CONDTIONAL PROBABILITY:

$$\boldsymbol{Pr(A \mid B)}$$

BAYES' THEOREM:

$$\boldsymbol{Pr(A|B) = \frac{Pr(A)Pr(B|A)}{Pr\,(B)}}$$

Bayesian techniques have been successfully used in a variety of applications - their use is no mere academic exercise in fancy statistics. They have *skin in the game*.

GALORATH

EXAMPLE 1: DRUG TESTING

LAW OF TOTAL PROBABILITY:

$$\mathbf{Pr(B) = Pr(}\boldsymbol{B}\mathbf{|}\boldsymbol{A}\mathbf{)\,Pr(}\boldsymbol{A}\mathbf{) + Pr(}\boldsymbol{B}\mathbf{|}\boldsymbol{A}^{'}\mathbf{)\,Pr(}\boldsymbol{A}^{'}\mathbf{)}$$

BAYES' THEOREM:

$$\mathbf{Pr(}\boldsymbol{A}\mathbf{|}\boldsymbol{B}\mathbf{) = \frac{Pr(}\boldsymbol{B}\mathbf{|}\boldsymbol{A}\mathbf{)\,Pr\,(}\boldsymbol{A}\mathbf{)}}{\mathbf{Pr(}\boldsymbol{B}\mathbf{|}\boldsymbol{A}\mathbf{)\,Pr(}\boldsymbol{A}\mathbf{) + Pr(}\boldsymbol{B}\mathbf{|}\boldsymbol{A}^{'}\mathbf{)\,Pr(}\boldsymbol{A}^{'}\mathbf{)}}$$

ANSWER:

$$\mathbf{Pr(}\boldsymbol{A}\mathbf{|}\boldsymbol{B}\mathbf{) = \frac{0.02(0.95)}{0.02(0.95) + 0.99(0.05)} \approx 27.7\%}$$

1

QUESTION

What is the probability that someone who fails a drug test is not a drug user?

2

ASSUMPTIONS

95% of the population are non-users

If someone is a drug user, it returns a positive result 99% of the time

If someone is not a user, it returns a positive result 2% of the time

3

DEFINING TERMS

A = Event that someone is a drug user

B = Event that someone test positive for drugs

4

COMPLEMENT

A' = Event that someone does NOT use drugs

B' = Event that someone tests negative for drugs

EXAMPLE 2: MONTY HALL PROBLEM

BAYES' THEOREM:

$$Pr(A_1|B) = \frac{Pr\,(A_1)\,Pr(B|A_1)}{Pr(A_1)\,Pr(B|A_1) + Pr(A_2)\,Pr(B|A_2) + Pr(A_3)\,Pr(B|A_3)}$$

POSTERIOR PROBABILITIES:

$$Pr(A_1|B) = \frac{(1/3)(1/2)}{(1/3)(1/2) + (1/3)(1) + (1/3)(0)} = \frac{1/6}{1/6 + 1/3} = 1/3$$

$$Pr(A_2|B) = \frac{(1/3)(1)}{(1/3)(1/2) + (1/3)(1) + (1/3)(0)} = \frac{1/3}{1/6 + 1/3} = 2/3$$

$$Pr(A_3|B) = 0$$

ANSWER: YOU SHOULD SWITCH!

1

QUESTION

There are three doors. Behind one is a car. Behind the other two cars are goats. You pick a door. Monty opens a different door and shows you a goat, and offers you the chance to switch

2

DEFINING TERMS

Let A_i denote the event that the car is behind the i^{th} door, WLOG assume:

1. You pick door #1
2. You are shown a goat behind door #2

You are shown a goat behind door #3, define this as event B

3

PRIOR PROBABILITY

Initial assumption is that

$P(A_1) = P(A_2) = P(A_3) = 1/3$

4

CONDITIONAL PROBABILITY

$P(B \mid A_3) = 0$

$P(B \mid A_2) = 1$

$P(B \mid A_1) = 1/2$

COST ANALYSIS

Applying Bayes' Theorem to Parametrics

RSDO EXAMPLE

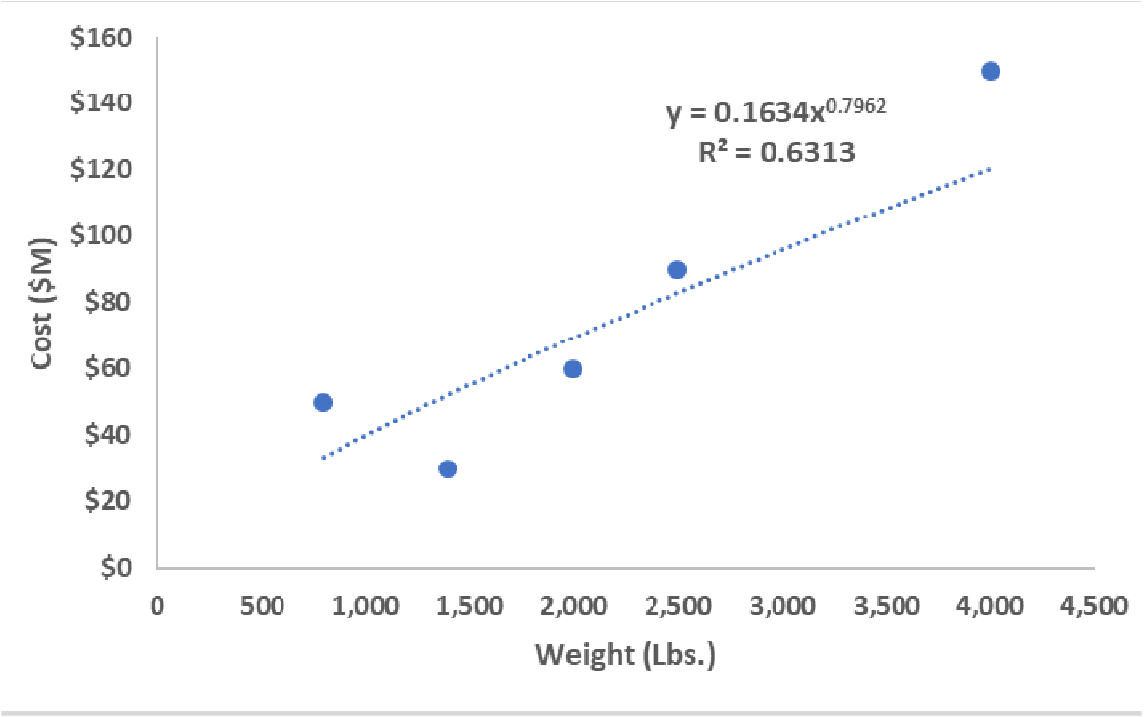
NASA's Rapid Spacecraft Development Office (RSDO) uses streamlined acquisition processes and fixed-price contracts to cut costs for robotic Earth-orbiting satellites

Give five historical data points of RSDO missions, what is the best way to estimate an RSDO mission

One option is to develop a CER with the five data points

However, given our discussion of the law of small numbers, a trend line developed from these five points may not be meaningful

Alternative – consider a larger set of robotic Earth-Orbiting satellites



GALORATH

COST ANALYSIS

Applying Bayes' Theorem to Parametrics

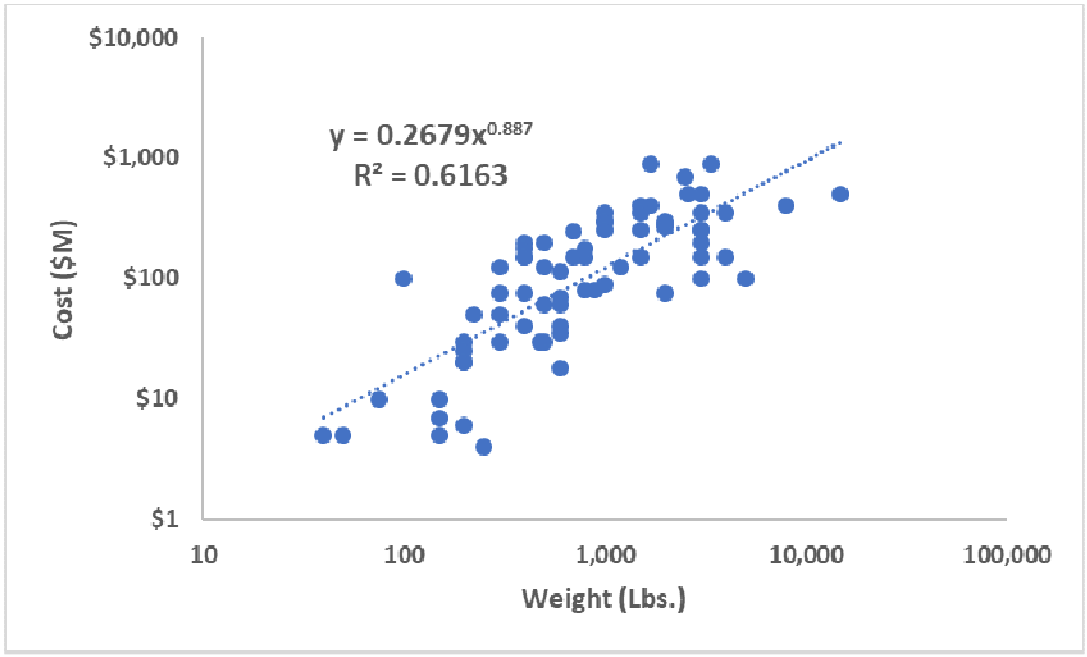
EARTH-ORBITING DATA

There are many more earth-orbiting data points if we do not restrict our attention to only RSDO missions

However, these missions are substantially more expensive than the RSDO analogues

Classical techniques apply here but they will likely overestimate a new RSDO mission by a significant amount

This is like the statistician who lost her car keys in a parking lot at night when someone saw her looking for them under a street light – when asked where she lost her keys she responded that the keys were near her car, far from the street light. When asked why she was looking near the street light, she responded that was where the light was shining



GALORATH

COST ANALYSIS

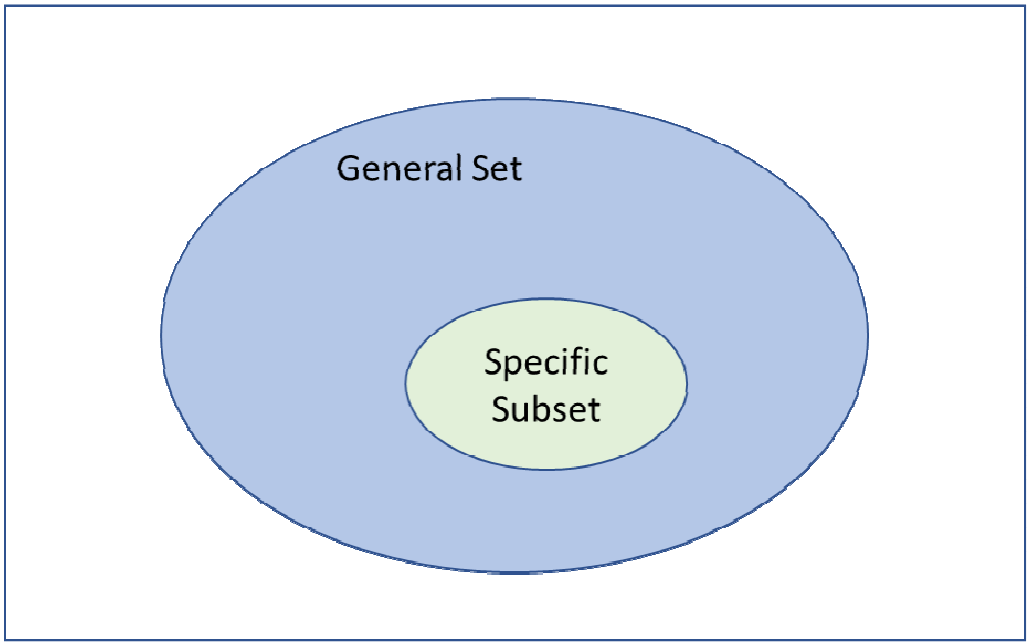
Applying Bayes' Theorem to Parametrics

COMBINING THE DATA

Bayes' Theorem provides a way to use both sources of data

The large, generic set of earth-orbiting missions can be used as a prior

The small specifically applicable data set can be used to update the prior probabilities



GALORATH

COST ANALYSIS

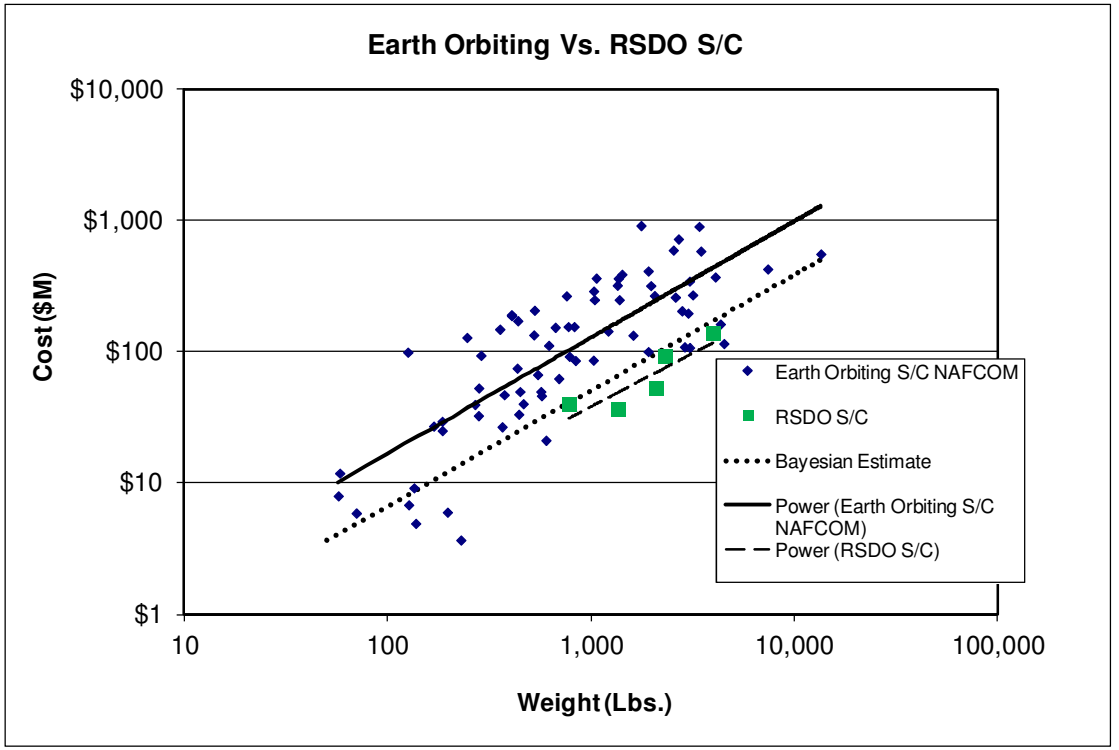
Applying Bayes' Theorem to Parametrics

RESULTS

The Bayesian CER coefficients are weighted averages of the coefficients of the CERs based on the two separate data sets

The Bayesian CER is applicable to a wider range of data than just the RSDO missions

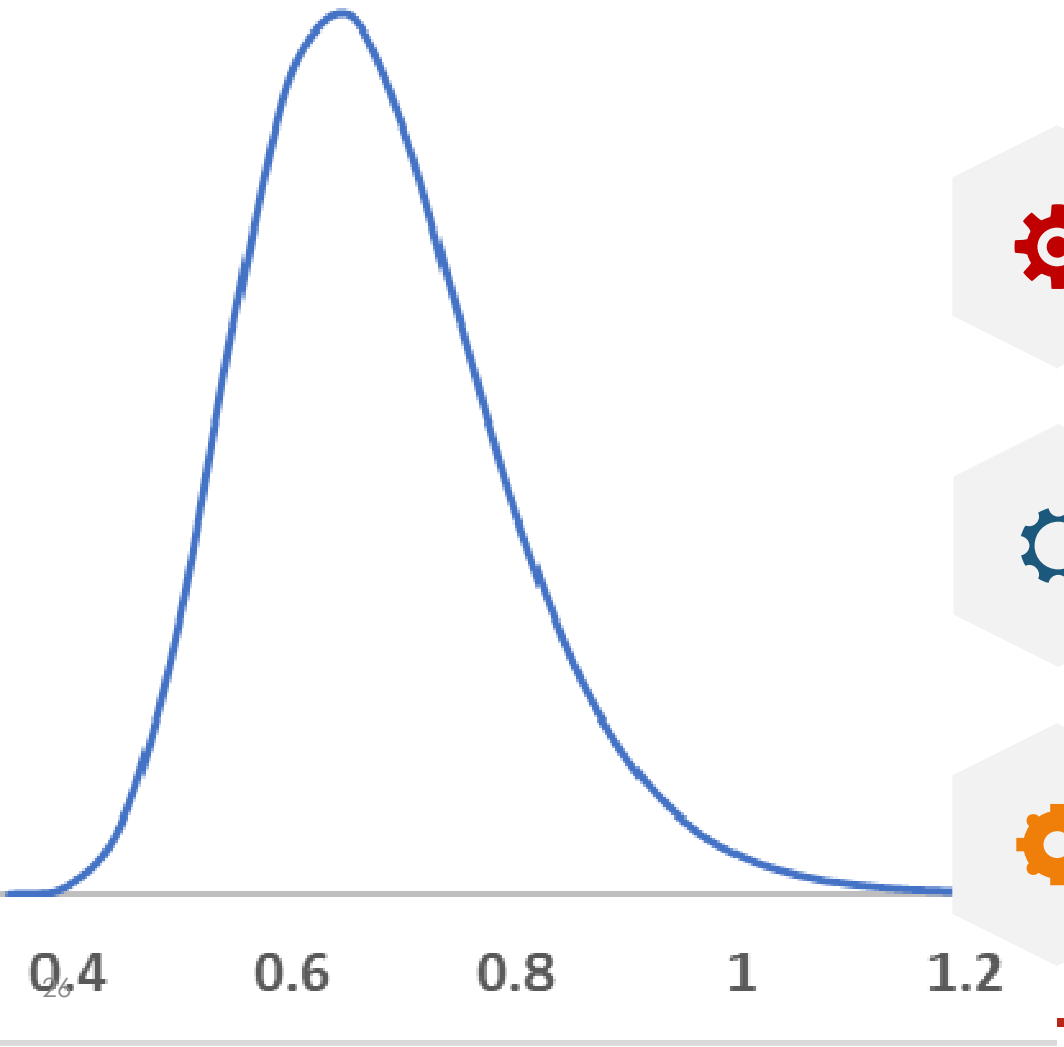
In testing the CER on an RSDO mission not in the data set, the Bayesian CER was more accurate than either of the RSDO-only CER or the all Earth-orbiting CER



GALORATH

THE BASIC METHOD AND ITS ISSUES

- **Everything we have discussed to this point relies on the use of a basic Bayesian method that has been presented before - for details see:**
  - **Christian Smart, "Bayesian Parametrics: How to Develop a CER with Limited Data and Even without Data," 2014 ICEAA Conference**
  - **Pierre Foussier, "The Benefits of the Bayesian Approach vs. the Frequentist Approach when Dealing with Low Data Sample," presented at the 2008 ISPA-SCEA international conference.**
- **There are some issues with this basic approach:**
  - **For a power equation, it requires the use of log-transformed ordinary least squares (biased low)**
  - **One assumption is that we know that the variance of the CER based on the sample data is known with certainty**
  - **Another assumption is that the residuals are lognormally distributed (normal in log space)**



UNKNOWN VARIANCE

THE MOST PROBLEMATIC ASSUMPTION IS THAT OF KNOW VARIANCE

In our example, the sample data is the small set, the one for which we have the least confidence in knowledge of the population variance

GOOD NEWS THIS CAN BE HANDLED ANALYITCALLY

By Cochran's Theorem, the variance can be modeled as a scaled inverse-Chi Square distribution

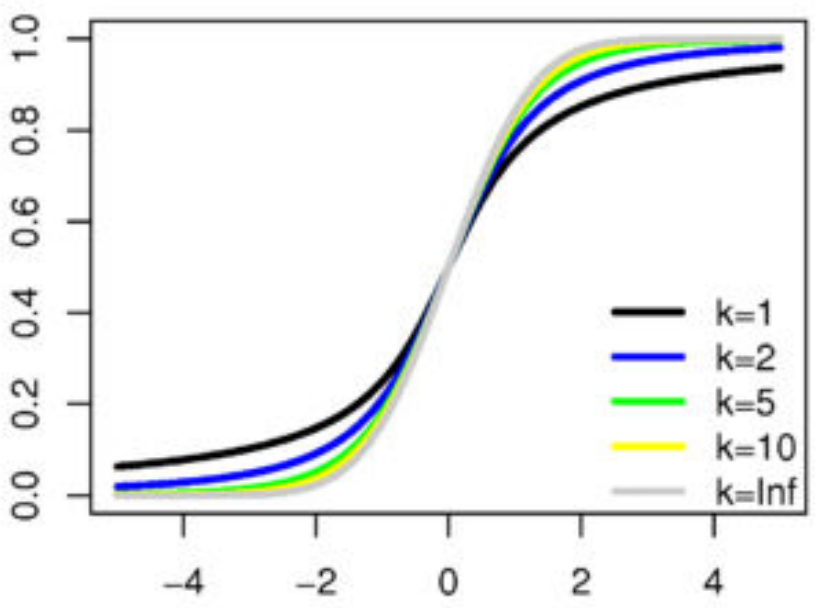
ESTIMATING PARAMETERS

Degrees of freedom is the same as the degrees of freedom in the regression of the sample data

Sample variance is the scaling factor

GALORATH

(LOG) NORMAL RESIDUALS



Student-t Distribution

GAUSSIAN ASSUMPTION MAKES THE MODEL ANALYTICALLY TRACTABLE

Gaussian likelihood has Gaussian as a conjugate prior

Even with unknown variance model is still tractable

Log-t has been proposed as an alternative to lognormal

Also - nonparametric methods have no distributional assumption

IF RESIDUALS ARE NOT GAUSSIAN, MUST USE SIMULATION

With small samples, residuals may be better modeled with log-t

Markov Chain Monte Carlo simulation can be used to do the Bayesian analysis via simulation

MARKOV CHAIN MONTE CARLO (MCMC)

MCMC uses conditional simulation – each trial depends on the previous

Can implement in R or use WinBUGS

GALORATH

RSDO - RESULTS COMPARISON

Confidence level	Normal likelihood with known variance	Normal likelihood with unknown variance	Student t likelihood, unknown variance, normal error	Student t likelihood, unknown variance, Student's t error
5%	$83	$44	$15	$23
10%	$94	$59	$25	$41
20%	$109	$83	$47	$72
30%	$121	$107	$76	$100
40%	$132	$133	$112	$129
50%	$144	$163	$163	$163
60%	$157	$199	$236	$205
70%	$171	$248	$352	$266
80%	$190	$319	$560	$370
90%	$220	$453	$1,068	$642
95%	$248	$606	$1,819	$1,169
99%	$311	$1,044	$4,942	$7,294
99.5%	$338	$1,275	$7,125	$21,659
99.9%	$401	$1,922	$15,151	$842,673

FOUR CASES CONSIDERED

1. Base case – Gaussian model (lognormal)
2. (Log) Gaussian residuals with unknown variance
3. Unknown variance, log-t for likelihood, lognormal predictive
4. Unknown variance, log-t for likelihood, log-t predictive

IMPLEMENTED IN BOTH R AND WINBUGS

Unknown variance changes the point estimate

The use of different uncertainty distributions does not change the point estimate (50th percentile)

THE DEVIL IS IN THE TAILS

the use of log-t dramatically effects the tails of the distribution

1 in 1,000 chance that a $163 million cost may grow to tens and hundreds of billions!

GALORATH

COMPARISON COMMENTS (1 OF 2)

- **Changing the assumption results in higher estimates**
  - **The 50th percentile for the Gaussian model is $144 million**
  - **Relaxing the variance assumption increases the 50th percentile to $163 million**
  - **Changing the CER residuals assumption does not change the 50th percentile**
  - **Improves the accuracy – actual cost was $180 million is 10% higher than the estimate with unknown variance**
- **There is also an increase in the heaviness of the right tail when relaxing the Gaussian assumptions**
  - **The unknown variance case has a 99th percentile equal to $1 billion**
  - **Extreme but it happens**

COMPARISON COMMENTS (2 OF 2)

- **When going to the log-t, the extreme right tail explodes, says that a relatively simple earth-orbiting spacecraft could cost as much as the James Webb Space Telescope (or more)!**
- **The relaxation of the assumption of known variance is important and the results are logical**
- **The change of the modeling of the residuals to a Student's t distribution in log space (log-t) is not logical**
- **Need to balance common sense with mathematical correctness – cost modeling is both an art and a science**

SUMMARY - BAYES

AN IDEAL METHOD WHEN YOU HAVE LIMITED DATA

MYRIAD REAL-WORLD APPLICATIONS

Cracking Enigma code

Search-and-Rescue

Property and Casualty Insurance Premium Setting

USES ALL YOUR DATA

Can be objective or subjective

Allows you to use all your data

FILLS A NEED

Ideal for small data sets

GAUSSIAN (BASIC) MODEL

Analytically tractable

Can be done in Excel

ADVANCED TECHNIQUES

The Gaussian model has issues with variance assumptions

If residuals are not lognormal, may require the use of Markov Chain Monte Carlo

LOW-HANGING FRUIT

Currently limited use

High potential

GALORATH

SUMMARY - OVERALL

HOW TO DEAL WITH SMALL DATA

ADMIT THERE IS AN ISSUE

Be aware that classical statistics does not work well for small data sets (<50 data points)

DO SOMETHING ABOUT IT

Collect data for more missions

Collect lower-level data

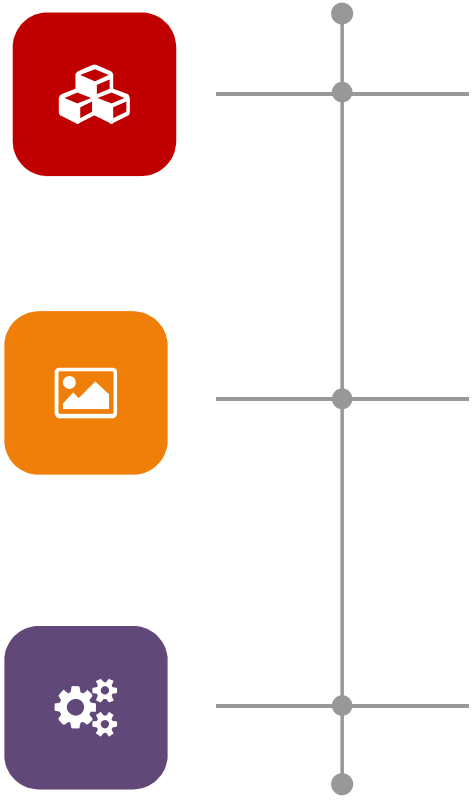
Imputation

Bayesian Methods

DATA COLLECTION

Time consuming and expensive

Low-hanging fruit has been collected (ONCE/CADRE)



IMPUTE

Use more of the data you already have by filling in gaps

Variety of statistical methods, well-developed

BAYES

Use Bayesian methods to leverage all your knowledge

LESSON

Avoid the abuse of large-sample methods on small data sets and use imputation and Bayesian methods

REFERENCES

- Boehm, B., A. Hira, K. Qi, and E. Venson, "Calibrating Use Case Points Using Bayesian Analysis," presented at the 2018 International Cost Estimating and Analysis Association Annual Conference.
- Cochran, W. G., 1934,"The distribution of quadratic forms in a normal system, with applications to the analysis of covariance," Mathematical Proceedings of the Cambridge Philosophical Society. 30 (2).
- Congdon, P., 2006, *Bayesian Statistical Modelling*, 2nd Edition, Wiley, West Sussex.
- Druker, E.R., R.L. Coleman, and P.J. Braxton, "Don't Let the Financial Crisis Happen to You: Why Estimates Using Power CERs are Likely to Experience Cost Growth," presented at the 2009 ISPA-SCEA Annual Conference.
- Foussier, P.M.M., "The Benefits of the Bayesian Approach vs. the Frequentist Approach when Dealing with Low Data Sample," presented at the 201o ISPA-SCEA conference.
- Smart, C.B., "Enhancing Risk Calibration Methods," presented at the 2018 International Cost Estimating and Analysis Association Annual Conference.
- Smart,C.B., "Cutting the Gordian Knot: Maximum Likelihood Estimation for Regression of Log Normal Error," presented at the 2017 International Cost Estimating and Analysis Association Annual Conference.
- Smart, C.B., "Covered with Oil: Incorporating Realism in Cost Risk Analysis," Journal of Cost Analysis and Parametrics, 2015.
- Smart, C.B., 2014, "Bayesian Parametrics: How to Develop a CER with Limited Data and Even Without Data," presented at the 2014 International Cost Estimating and Analysis Association Annual Conference.
- Taleb, N.N., *Skin in the Game: Hidden Asymmetries in Daily Life*, Random House, New York, 2018.

GALORATH